ABSTRAK

Salah satu aspek sebagai indikator kualitas sekolah menengah atas adalah tingkat diterimanya siswa di perguruan tinggi negeri. Beberapa data siswa sekolah menengah atas dianalisis untuk mengetahui tingkat diterimanya siswa di perguruan tinggi negeri. Proses analisis data siswa tersebut menggunakan teknik *data mining*. Tujuan penelitian ini adalah (1) mengetahui penerapan klasifikasi *naïve bayes* pada rata-rata nilai rapor, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua siswa terhadap tingkat diterimanya siswa di perguruan tinggi negeri dan (2) memprediksi siswa angkatan 2015 yang diterima di perguruan tinggi negeri menggunakan hasil model klasifikasi yang terbentuk.

Pada penelitian ini data yang digunakan adalah data siswa angkatan 2010, 2011, 2013 dan 2014 Sekolah Menengah Atas Negeri 1 Sleman yang berjumlah 807 data, setelah dilakukan *pre-processing* diperoleh data yang siap diolah berjumlah 716 data. Proses *data mining* dibantu oleh *software* WEKA menggunakan klasifikasi *naïve bayes* dan 10-*fold cross validation*.

Proses evaluasi data tersebut menghasilkan tingkat akurasi sebesar 63,83%, selanjutnya model klasifikasi *naïve bayes* digunakan untuk mengolah data prediksi. Data prediksi yang digunakan adalah data siswa angkatan 2015. Hasil prediksi dari 172 siswa adalah terdapat 14 siswa diterima di perguruan tinggi negeri dan sisanya tidak diterima di perguruan tinggi negeri.

Kata kunci: *data mining*, klasifikasi *naïve bayes*, perguruan tinggi negeri, data siswa, *software* WEKA

BAB I

PENDAHULUAN

A. Latar Belakang Masalah

Kualitas pendidikan pada siswa merupakan salah satu tujuan dari sekolah. Sekolah akan meningkatkan kualitasnya dari tahun ke tahun sebagai tolok ukur dalam menentukan keberhasilan sistem pendidikannya. Salah satu aspek sebagai indikator kualitas di sekolah adalah tingkat kelulusan siswa dan banyaknya lulusan siswa sekolah menengah atas yang diterima di perguruan tinggi negeri juga menunjukkan kualitas sekolah dalam memotivasi siswa untuk memberikan jaminan kesempatan kerja yang lebih baik di masa mendatang.

Setiap tahun, jumlah data siswa di sekolah bertambah. Data tersebut tersimpan dalam bentuk *hard file* misalnya hasil *print out* data, catatan buku dan dalam bentuk *soft file* pada komputer. Hal tersebut mengakibatkan menumpuknya data yang disimpan oleh sekolah. Keadaan seperti ini digambarkan sebagai sebuah situasi "*data rich but information poor*" yang artinya data yang disimpan berlimpah namun informasi yang diperoleh kurang. Bertambahnya jumlah data yang cepat dan besar yang dikumpulkan dan disimpan setiap tahunnya jauh melampaui kemampuan manusia dalam menganalisis data tanpa suatu teknik yang tepat. Sebagai akibatnya, data yang dikumpulkan menjadi sesuatu yang tidak bermanfaat. Untuk mengatasi hal tersebut diperlukan suatu teknik dalam *data mining* yang dapat mengubah sesuatu yang tidak bermanfaat menjadi sesuatu yang berharga dan dapat memberikan suatu informasi yang penting. Salah satu sekolah yang mengalami hal tersebut adalah Sekolah Menengah Atas Negeri 1 Sleman.

Sekolah Menengah Atas Negeri 1 Sleman termasuk sekolah unggulan di Kabupaten Sleman. Banyak pula siswa yang lulus dan diterima pada seleksi perguruan tinggi negeri melalui jalur seleksi tulis maupun seleksi non tulis, namun, ada pula siswa yang tidak meneruskan sekolahnya ke jenjang yang lebih tinggi dikarenakan berbagai hal.

Keputusan seleksi masuk perguruan tinggi negeri ditentukan oleh beberapa faktor. Salah satu faktor penentu kecermatan prediksi adalah pemilihan prediktor, dimana semakin tepat pemilihan prediktor, maka akan semakin tepat pula pengambilan keputusan dalam seleksi (Nandan Supriatna, 2009).

Menurut Nandan Supriatna (2009), model seleksi penerimaan mahasiswa menggunakan rata-rata nilai rapor sekolah menengah atas. Penggunaan nilai rapor akan memungkinkan pemilihan prediktor seleksi yang sesuai dengan kebutuhan setiap program studi yang ada di perguruan tinggi negeri, sehingga diharapkan akan mempunyai daya prediksi yang tinggi. Disamping itu, nilai ujian nasional merupakan informasi tambahan yang digunakan sebagai bahan pertimbangan dalam penentuan penerimaan calon mahasiswa. Setelah tiga tahun siswa bersekolah, maka mereka akan menempuh ujian nasional yang merupakan sesuatu yang diwajibkan bagi para siswa untuk persyaratan kelulusan (Faiz Hidayat, 2013). Selain menjadi syarat kelulusan, nilai ujian nasional juga digunakan sebagai faktor yang berpengaruh terhadap diterimanya siswa di perguruan tinggi negeri.

Tingkat pendidikan orang tua akan menentukan cara orang tua dalam membimbing dan mengarahkan anaknya dalam hal pendidikan yang sedang maupun yang akan ditempuh anaknya. Orang tua yang berwawasan luas akan

mengarahkan dan membimbing anaknya untuk terus menimba ilmu melebihi orang tuanya. Semakin tinggi tingkat pendidikan orang tua maka semakin tinggi pula minat siswa untuk melanjutkan sekolah ke perguruan tinggi (Esti Setya R, 2012). Hal inilah yang menjadi latar belakang tingkat pendidikan orang tua menjadi salah satu prediktor yang mempengaruhi siswa untuk melanjutkan sekolahnya ke perguruan tinggi negeri.

Jenis pekerjaan orang tua berpengaruh dalam tingkat sosial ekonomi keluarga. Semakin baik jenis pekerjaan orang tua maka semakin tinggi tingkat pendapatan orang tua. Pendapatan orang tua dapat menunjang pendidikan anaknya dalam bentuk materi, baik dalam memberi fasilitas untuk kegiatan pembelajaran, biaya sekolah, dan sebagainya, sehingga jenis pekerjaan orang tua menjadi salah satu prediktor yang berpengaruh dalam tingkat studi siswa selanjutnya.

Menurut Ayinde, et al (2013), tidak hanya bermacam-macam faktor seperti kepribadian siswa, keadaan sosial ekonomi, psikologi, dan variabel dari lingkungan lainnya yang mempengaruhi pendidikan siswa, namun juga model yang digunakan untuk memprediksi data siswa yang diperoleh dari literatur dan beberapa studi yang spesifik.

Proses menganalisis data lebih lanjut diperlukan untuk mengetahui apakah prediktor di atas berpengaruh pada tingkat diterimanya siswa sekolah menengah atas di perguruan tinggi negeri. Proses tersebut menggunakan teknik *data mining*, dengan teknik tersebut dapat diketahui seberapa besar tingkat diterimanya siswa berdasarkan rata-rata nilai rapor, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua. Proses di dalam *data mining* untuk menemukan informasi pada

dan kecepatan yang tinggi saat diaplikasikan untuk jumlah data yang besar. Klasifikasi *naïve bayes* adalah salah satu teknik *data mining* yang paling populer untuk mengklasifikasikan data dalam jumlah yang besar dan dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. Hal tersebut dapat diterapkan pada masalah klasifikasi seperti peramalan cuaca, deteksi gangguan, diagnosis penyakit, dan lain-lain (Kabir, et al, 2011).

Berdasarkan uraian di atas, maka akan dilakukan analisis terhadap rata-rata nilai rapor dari semester satu sampai semester lima, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua siswa Sekolah Menengah Atas Negeri 1 Sleman untuk memperoleh informasi siswa yang diterima di perguruan tinggi negeri menggunakan klasifikasi *naïve bayes*. Data tersebut diolah sedemikian rupa sehingga membentuk sebuah model klasifikasi, sehingga model klasifikasi tersebut dapat digunakan untuk mengetahui tingkat diterimanya siswa di perguruan tinggi negeri untuk angkatan selanjutnya.

B. Perumusan Masalah

Perumusan masalah pada penelitian ini adalah sebagai berikut

1. Bagaimana penerapan klasifikasi *naïve bayes* pada rata-rata nilai rapor, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua siswa Sekolah Menengah Atas Negeri 1 Sleman terhadap tingkat diterimanya siswa di perguruan tinggi negeri berdasarkan data siswa angkatan 2010, 2011, 2013, dan 2014?
2. Bagaimana hasil prediksi siswa angkatan 2015 Sekolah Menengah Atas Negeri 1 Sleman yang diterima di perguruan tinggi negeri menggunakan hasil model klasifikasi *naïve bayes* yang terbentuk?

C. Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut

1. menerapkan klasifikasi *naïve bayes* pada rata-rata nilai rapor, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua siswa Sekolah Menengah Atas Negeri 1 Sleman terhadap tingkat diterimanya siswa di perguruan tinggi negeri berdasarkan data siswa angkatan 2010, 2011, 2013, dan 2014,
2. memprediksi siswa angkatan 2015 Sekolah Menengah Atas Negeri 1 Sleman yang diterima di perguruan tinggi negeri menggunakan hasil model klasifikasi *naïve bayes* yang terbentuk.

D. Manfaat Penelitian

Berdasarkan berbagai informasi yang diperoleh, penyusunan tugas akhir ini diharapkan dapat memberikan manfaat-manfaat sebagai berikut

1. memperoleh suatu informasi terkait tingkat diterimanya siswa di perguruan tinggi negeri sehingga dapat digunakan sebagai patokan untuk meningkatkan kualitas pendidikan di sekolah,
2. menambah ilmu pengetahuan yang dapat dikembangkan ke tingkat lebih lanjut.

E. Batasan Masalah

Batasan permasalahan dalam penelitian ini adalah sebagai berikut

1. penelitian ini hanya meneliti tentang tingkat diterimanya siswa Sekolah Menengah Atas Negeri 1 Sleman di perguruan tinggi negeri berdasarkan rata-rata nilai rapor, nilai ujian nasional, pendidikan orang tua dan pekerjaan orang tua,
2. data yang digunakan adalah data siswa tahun 2010, 2011, 2013, dan 2014 yang diperoleh dari Bagian Tata Usaha, Bagian Kurikulum dan Bagian Bimbingan Konseling Sekolah Menengah Atas Negeri 1 Sleman.

BAB II

KAJIAN TEORI

Bab ini akan membahas tentang pengertian dasar basis data, *data mining*, klasifikasi, *naïve bayes*, faktor-faktor yang berpengaruh dalam diterimanya siswa di perguruan tinggi negeri, perguruan tinggi negeri, *software* WEKA, *k-fold cross validation*, dan penelitian yang relevan.

A. *Database*

Sebuah sistem *database* atau yang disebut *Database Management System* (*DBMS*) adalah sekumpulan data yang saling berhubungan yang dikenal sebagai basis data dan sekumpulan program perangkat lunak untuk mengatur dan mengakses data (Han, et al, 2012: 9). Tujuan utama DBMS adalah untuk menyimpan dan memberikan informasi pada *database* dengan tepat dan efisien (Silberschatz, et al, 2006: 1). Menurut Han, et al (2012: 9) jenis-jenis *database* adalah sebagai berikut

1. *Relational database*

   Relational database atau basis data relasional adalah sebuah kumpulan tabel dengan nama khusus dan setiap tabel terdiri atas kumpulan atribut (kolom atau *field*) dan biasanya menyimpan data dalam jumah yang besar pada data (baris atau *record*). Setiap data dalam tabel relasi menunjukkan sebuah objek yang diindentifikasi oleh sebuah *unique key* dan digambarkan oleh nilai dari atribut tersebut.

lengkap, menganalisis data, perubahan bentuk data, penulisan ulang data-data yang baru, dan pembaharuan data secara periodik.

3. *Transactional Data*

   Transactional data pada setiap *record* dikumpulkan berdasarkan sebuah transaksi (dalam dunia bisnis). Sebuah transaksi memliki nomor identitas transaksi yang unik (*trans_ID*). *Transactional data* yang mempunyai tabel tambahan yang berisi informasi lain direlasikan pada hubungan yang mungkin terjadi, seperti deskripsi barang, informasi dari pelayan toko, dan lain-lain.

Sistem *database* dirancang untuk mengatur informasi dalam jumlah yang besar sehingga sistem *database* memberikan keamanan untuk menyimpan data, meskipun terjadi kerusakan pada sistem.

Database yang digunakan dalam penelitian ini adalah data siswa tahun 2010, 2011, 2013, dan 2014 Sekolah Menengah Atas Negeri 1 Sleman yang berupa data induk siswa dan data kelulusan siswa, sedangkan nomor identitas siswa (NIS) digunakan sebagai *primary key* untuk menghubungkan data siswa dengan data kelulusan siswa.

B. *Data Mining*

Data mining adalah proses menentukan pola dan informasi dari data yang berjumlah besar. Sumber data dapat berupa *database*, *data warehouse*, *Web*, tempat penyimpanan informasi lainnya atau data yang mengalir ke dalam sistem yang dinamis (Han, et al, 2012: 8).

Menurut Gartner Group, *data mining* adalah suatu proses menemukan hubungan baru, pola dan kecenderungan dengan penyaringan pada kumpulan data yang besar yang tersimpan dalam tempat penyimpanan, menggunakan teknologi pengenalan pola seperti teknik statistik dan teknik matematika (Larose, 2005: 2).

Data mining mempunyai beberapa metode yang dilakukan pengguna untuk meningkatkan proses *mining* supaya lebih efektif. Oleh karena itu, data mining dibagi menjadi beberapa kelompok berdasarkan metodenya, yaitu (Larose, 2005: 11-17)

1. Estimasi

   Variabel target pada proses estimasi lebih condong ke arah numerik daripada ke arah kategorikal (nominal). Model dibangun menggunakan *record* lengkap yang menyediakan nilai dari variabel target sebagai nilai prediksi, kemudian nilai estimasi dari variabel target dibuat berdasarkan pada nilai prediksi. Contoh algoritma estimasi adalah *Linear Regression* dan *Neural Network*.

2. Prediksi

   Prediksi hampir sama dengan klasifikasi dan estimasi, namun, pada prediksi data yang digunakan adalah data rentet waktu (*data time series*) dan

hasil akhir nilainya digunakan untuk beberapa waktu mendatang. Contoh prediksi dalam bisnis dan penelitian adalah prediksi harga beras dalam tiga bulan kedepan.

3. Klasifikasi

   Pada klasifikasi terdapat variabel target yang berupa nilai kategorikal (nominal). Contoh dari klasifikasi adalah pendapatan masyarakat digolongkan ke dalam tiga kelompok, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah. Algoritma klasifikasi yang biasa digunakan adalah *Naïve Bayes*, *K-Nearest Neighbor*, dan *C4.5*.

4. Klastering

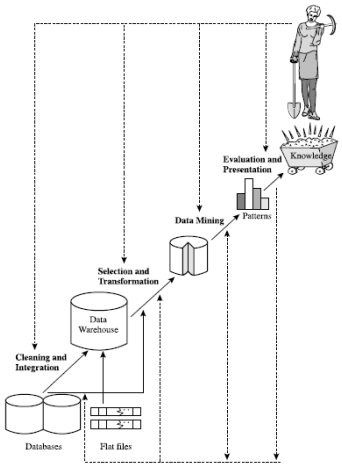
   Klastering merupakan pengelompokan data atau pembentukan data ke dalam jenis yang sama. Klastering tidak untuk mengklasifikasi, mengestimasi, atau memprediksi nilai, tetapi membagi seluruh data menjadi kelompok-kelompok yang relatif sama (homogen). Perbedaan algoritma klastering dengan algoritma klasifikasi adalah klastering tidak memiliki *target*/*class*/*label*, jadi klastering termasuk *unsupervised learning*. Contoh algoritma klastering adalah *K-Means* dan *Fuzzy C-Means*.

5. Aturan Asosiasi

   Aturan asosiasi digunakan untuk menemukan atribut yang muncul dalam waktu yang bersamaan dan untuk mencari hubungan antara dua atau lebih data dalam sekumpulan data. Contoh penggunaan aturan asosiasi adalah analisis kemungkinan seorang pelanggan membeli roti dan susu dalam waktu

yang bersamaan di suatu pasar swalayan. Contoh algoritma aturan asosiasi yang sering digunakan adalah *Apriori* dan *FP-Growth*.

Istilah *data mining* dan *knowledge discovery in databases* (KDD) sering digunakan untuk menjelaskan proses pencarian informasi yang tersembunyi di dalam suatu *database* yang besar. Sebelum melakukan proses data mining, diperlukan proses KDD terlebih dahulu supaya menghasikan kualitas data yang baik. Ilustrasi dari proses pada KDD dijelaskan pada Gambar 2.1 berikut :



Gambar 2.1 Proses dalam KDD (Han, et al, 2012: 7)

ini bertujuan untuk membersihkan data dari data yang tidak konsisten. Biasanya, data yang diperoleh ada yang tidak lengkap, seperti data yang hilang, salah ketik, dan sebagainya. Data-data tersebut lebih baik dibuang karena akan mempengaruhi kinerja proses selanjutnya.

2. *Data integration*

   Data integration adalah proses untuk menggabungkan beberapa sumber data ke dalam satu *database*. Proses ini perlu dilakukan dengan cermat dan teliti karena apabila terjadi kesalahan pada integrasi data maka bisa memberikan hasil yang menyimpang dan mempengaruhi proses selanjutnya.

3. *Data selection*

   Data selection adalah proses menganalisis data-data yang relevan dari *database* karena sering ditemukan bahwa tidak semua data dibutuhkan dalam proses *data mining*. Data tersebut dipilih dan diseleksi dari *database* untuk dianalisis.

4. *Data transformation*

   Tipe data pada *database* diubah ke dalam pola tertentu sehingga dapat diproses dalam *data mining*. Proses ini sangat bergantung pada jenis yang dibutuhkan dalam *database*.

5. *Data mining*

   Data mining adalah proses pokok dalam KDD. Proses ini menggunakan metode yang tepat dengan tujuan untuk menghasilkan pola data tertentu. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

6. *Pattern evaluation*

   Pattern evaluation adalah proses untuk mengidentifikasi pola yang tepat. Pola-pola tersebut dievaluasi untuk menilai apakah hasil yang diharapkan tercapai atau tidak. Jika hasil yang diperoleh tidak sesuai maka terdapat beberapa cara untuk memperbaiki hal tersebut, salah satu contohnya yaitu mencoba metode *data mining* yang lain.

7. *Knowledge presentation*

   Proses ini merupakan visualisasi dan representasi teknik yang digunakan untuk memperoleh informasi sehingga informasi yang dihasilkan mudah dimengerti oleh orang-orang yang tidak memahami *data mining*.

C. Klasifikasi

Proses untuk menemukan pola yang menjelaskan data yang penting dikenal sebagai klasifikasi. Metode klasifikasi dalam data mining ada banyak, diantaranya *decision tree*, *k-nearest neighbours*, *neural network* dan *naïve bayes*.

1. *Decision Tree*

   Decision tree merupakan metode klasifikasi dalam bentuk diagram yang direpresentasikan seperti struktur pohon, setiap simpul internal

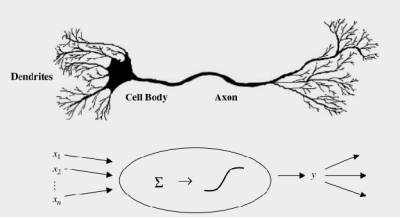
merupakan sebuah atribut, setiap cabang merupakan nilai atribut, dan setiap simpul daun atau simpul terminal merupakan label *class*, serta simpul yang paling atas adalah simpul akar (Han, et al, 2012: 330). Metode ini popular karena model yang terbentuk mudah dipahami. Salah satu kekurangan pada *decision tree* adalah membutuhkan waktu dan jumlah memori yang banyak untuk data yang besar dalam mendesain pohon keputusan yang optimal.

2. *K-Nearest Neighbor*

   Metode *k-nearest neighbor* merupakan metode klasifikasi pertama yang dijabarkan pada awal tahun 1950. Metode ini sulit digunakan pada data dalam jumlah yang besar sehingga tidak terkenal sampai tahun 1960 ketika kemampuan teknologi meningkat metode ini mulai dipakai (Han, et al, 2012: 423). *K-nearest neighbor* adalah metode klasifikasi untuk menghitung kedekatan antara atribut baru dengan atribut lama berdasarkan bobot setiap atribut tersebut (Kusrini dan Emha, 2009: 93). Metode ini membutuhkan waktu unuk menentukan *k* (jumlah tetangga terdekat) yang bernilai optimal, namun memberikan hasil evaluasi data lebih akurat dan cocok untuk data yang besar.

3. *Neural Network*

   Neural nerwork terinspirasi oleh pengenalan sistem pembelajaran yang kompleks pada otak binatang yang terdiri atas kumpulan neuron yang saling berhubungan (Larose, 2005: 128). Gambar 2.2 berikut adalah ilustrasi kinerja *neural network* yang meniru kinerja neuron asli dalam otak.



Gambar 2.2 Neuron Asli dan Model Neuron Buatan (Larose, 2005: 129)

Salah satu kelebihan *neural network* adalah tanggap dengan data *noisy*, karena jaringan ini terdiri dari banyak *node* dan setiap sambungan *node* mempunyai tugas masing-masing sehingga jaringan dapat bekerja maksimal pada data yang tidak berisi keterangan atau pada data yang *error*, namun kekurangannya, *neural nerwork* membutuhkan waktu yang lama dalam kinerjanya dan orang awam susah untuk mengartikan simbol pada bobot dan unit di jaringannya (Larose, 2005: 129).

4. *Naïve Bayes*

   Klasifikasi *naïve bayes* adalah salah satu teknik *data mining* yang paling populer untuk mengklasifikasikan data dalam jumlah yang besar dan dapat digunakan untuk memprediksi probabilitas keanggotaan suatu *class*. Hal tersebut dapat diterapkan pada masalah klasifikasi seperti peramalan cuaca, deteksi gangguan, diagnosis penyakit, dll (Kabir, et al, 2011). *Naïve bayes* dipilih karena memiliki tingkat ketelitian dan kecepatan yang tinggi saat diaplikasikan untuk jumlah data yang besar.

Benevolence, Or An Attempt to Prove that the Principal End of the Divine Providence and Government is the Happiness of His Creatures" yang dikenal sebagai perintis probabilitas (Bramer, 2009: 24).

Teorema *bayes* memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya. Pada teorema *Bayes, X* dijabarkan oleh kumpulan *n* atribut dengan H adalah beberapa hipotesis, sehingga data *X* termasuk sebuah *class C* (Han, et al, 2012: 350). Dengan teorema *bayes,* sebagai berikut

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2.1)$$

Klasifikasi *naïve bayes* yang mengacu pada teorema *bayes* mempunyai persamaan sebagai berikut

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (2.2)$$

Menurut Han, et al (2012: 351) proses dari pengklasifikasian *naïve bayes* adalah sebagai berikut

1. variabel *D* adalah kumpulan dari data dan label yang terkait dengan *class*. Setiap data diwakili oleh vektor atribut n-dimensi, $X = (x_1, x_2, \dots, x_n)$ dengan n dibuat dari data n atribut, berturut-turut, $A_1, A_2, \dots, A_n$.
2. misalkan terdapat i *class*, $C_1, C_2, \dots, C_i$. Diberikan sebuah data *X*, kemudian pengklasifikasian akan memprediksi *X* ke dalam kelompok yang memiliki

probabilitas posterior tertinggi berdasarkan kondisi *X*. Artinya, pengklasifikasian *naïve bayes* memprediksi bahwa data *X* termasuk *class* C_i, jika dan hanya jika

$$P(C_i|X) > P(C_j|X) \text{ untuk } 1 \le j \le m, j \ne i \quad (2.3)$$

maka nilai $P(C_i|X)$ harus lebih dari nilai $P(C_j|X)$ supaya diperoleh hasil akhir $P(C_i|X)$.

3. ketika $P(X)$ konstan untuk semua *class* maka hanya $P(X|C_i)\,P(C_i)$ yang dihitung. Jika probabilitas *class prior* sebelumnya tidak diketahui, maka diasumsikan bahwa *class*-nya sama, yaitu $P(C_1) = P(C_2) = \cdots = P(C_m)$, untuk menghitung $P(X|C_i)$ dan $P(X|C_i)\,P(C_i)$. Perhatikan bahwa probabilitas *class prior* dapat diperkirakan oleh

$$P(C_i) = \frac{(|C_{(i,D)}|)}{|D|} \quad (2.4)$$

dimana $|C_{i,D}|$ adalah jumlah data *training* dari *class* C_i dan *D* adalah jumlah total data *training* yang digunakan.

4. apabila diberikan kumpulan data yang mempunyai banyak atribut, maka mengurangi perhitungan $P(X|C_i)$, *naïve bayes* mengasumsikan pembuatan *class independen* yang bersyarat. Anggap bahwa nilai-nilai atribut tersebut bersifat independen satu sama lain dan diantara atribut tidak terdapat relasi depedensi, maka

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \dots \times P(x_n|C_i) \quad (2.5)$$

Perhitungan $P(X|C_i)$ pada setiap atribut mengikuti hal-hal berikut

a. jika A_k adalah kategorikal, maka $P(x_k|C_i)$ adalah jumlah data dari *class* C_i di *D* yang memiliki nilai x_k untuk atribut A_k dibagi dengan $|C_{i,D}|$ yaitu jumlah data dari *class* C_i di *D,*
b. jika A_k adalah numerik, biasanya diasumsikan memiliki distribusi *Gauss* dengan rata-rata μ dan standar deviasi σ, didefinisikan oleh

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (2.6)$$

sehingga diperoleh

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (2.7)$$

Setelah itu akan dihitung μ_{C_i} dan σ_{C_i} yang merupakan deviasi *mean* (rata-rata) dan standar deviasi masing-masing nilai atribut A_k untuk *training tuple class* C_i.

5. $P(X|C_i)P(C_i)$ dievaluasi pada setiap *class* C_i untuk memprediksi pengklasifikasian label *class* data *X* dengan menggunakan

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \text{ untuk } 1 \le j \le m,\ j \ne i \quad (2.8)$$

label *class* untuk data *X* yang diprediksi adalah *class* C_i jika nilai $P(X|C_i)P(C_i)$ lebih dari nilai $P(X|C_j)P(C_j)$.

Sebagai contoh penerapan dari *naïve bayes*, diberikan perhitungan manual pada sebuah kasus. Tabel 2.1 terdapat *data training* dari *database* pelanggan *AllElectronics*.

Tabel 2.1 *Database* Pelanggan *AllElectronics*

id	*age*	*Income*	*student*	*credit_rating*	*buys_computer*
1	≤ 30	*High*	*No*	*Fair*	*No*
2	≤ 30	*High*	*No*	*Excellent*	*No*
3	31 - 40	*High*	*No*	*Fair*	*Yes*
4	> 40	*Medium*	*No*	*Fair*	*Yes*
5	> 40	*Low*	*Yes*	*Fair*	*Yes*
6	> 40	*Low*	*Yes*	*Excellent*	*No*
7	31 - 40	*Low*	*Yes*	*Excellent*	*Yes*
8	≤ 30	*Medium*	*No*	*Fair*	*No*
9	≤ 30	*Low*	*Yes*	*Fair*	*Yes*
10	> 40	*Medium*	*Yes*	*Fair*	*Yes*
11	≤ 30	*Medium*	*Yes*	*Excellent*	*Yes*
12	31 - 40	*Medium*	*No*	*Excellent*	*Yes*
13	31 - 40	*High*	*Yes*	*Fair*	*Yes*
14	> 40	*Medium*	*No*	*Excellent*	*No*

Berdasarkan Tabel 2.1, terdapat dua *class* target yaitu pelanggan membeli computer atau "*buys_computer= yes"* dan pelanggan tidak membeli computer atau "*buys_computer=no"*

Misalkan diberikan sebuah contoh data yang belum diketahui *class* targetnya. Contoh tersebut diberikan pada Tabel 2.2 yaitu

Tabel 2.2 Data Prediksi

id	*age*	*Income*	*student*	*credit_rating*	*buys_computer*
15	≤ 30	*Medium*	*Yes*	*Fair*	

pada data prediksi tersebut dengan $P(C_i)$ merupakan *class target*, kemudian akan ditentukan *class* atribut yang digunakan dengan ketentuan

C_1 = (*buys_computer*="*yes*")

C_2 = (*buys_computer*="*no*")

x_1 = (*age*="≤ 30")

x_2 = (*income*="*medium*")

x_3 = (*student*="*yes*")

x_4 = (*credit_rating*="*fair*")

Berikut adalah langkah dalam perhitungan *naïve bayes* dengan menghitung probabilitas *class* C_i untuk $i = 1, 2$ berdasarkan jumlah nilai atribut pada Tabel 2.1

1. pada *class* "*buys_computer*" yang benilai "*yes*" sebanyak sembilan data dan *class* "*buys_computer*" yang benilai "*no*" sebanyak lima data. Maka akan dihitung nilai $P(C_i)$, yaitu

   $P(C_1)$ = P(*buys_computer*="*yes*") = $\frac{9}{14} = 0{,}643$

   $P(C_2)$ = P(*buys_computer*="*no*") = $\frac{5}{14} = 0{,}357$

2. menghitung nilai $P(X|C_i)$ untuk $i = 1, 2$, 1 = *yes*, 2 = *no*, yaitu

   $P(x_1|C_1)$ = P(*age*="≤ 30" | *buys_computer*="*yes*") = $\frac{2}{5} = 0{,}4$

   $P(x_1|C_2)$ = P(*age*="≤ 30" | *buys_computer*="*no*") = $\frac{3}{5} = 0{,}600$

$P(x_2|C_1)$ = P(*income*="*medium*" | *buys_computer*="*yes*") = $\frac{4}{9} = 0{,}444$

$P(x_2|C_2)$ = P(*income*="*medium*" | *buys_computer*="*no*") = $\frac{2}{5} = 0{,}400$

$P(x_3|C_1)$ = P(*student*="*yes*" | *buys_computer*="*yes*") = $\frac{6}{9} = 0{,}667$

$P(x_3|C_2)$ = P(*student*="*yes*" | *buys_computer*="*no*") = $\frac{1}{5} = 0{,}200$

$P(x_4|C_1)$ = P(*credit_rating*="*fair*" | *buys_computer*="*yes*") = $\frac{6}{9} = 0{,}667$

$P(x_4|C_2)$ = P(*credit_rating*="*fair*" | *buys_computer*="*no*") = $\frac{2}{5} = 0{,}400$

3. menghitung nilai $P(X|C_i)P(C_i)$ untuk $i = 1, 2$ maka digunakan rumus (2.4), yaitu

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(x_n|C_i)$$

Perhitungan untuk $i = 1$ bernilai "*yes*"

$P(X \mid buys\_computer="yes")$ = P(*age*="≤ 30"| *buys_computer*="*yes*") × P(*income*="*medium*"| *buys_computer*="*yes*") × P(*student*="*yes*"| *buys_computer* ="*yes*") × P(*credit_rating*="*fair*" | *buys_computer*="*yes*")

= 0,079

Perhitungan untuk $i = 2$ bernilai "*no*"

$P(X \mid buys\_computer="no")$ = P(*age*="≤ 30" | *buys_computer*="*no*") × P(*income*="*medium*" | *buys_computer*="*no*") × P(*student*="*yes*" | *buys_computer*="*no*") × P(*credit_rating*="*fair*" | *buys_computer*="*no*")

= 0,019

4. selanjutnya menghitung nilai $P(X|C_i)P(C_i)$ yaitu

   $P(X \mid buys\_computer="yes")\ P(buys\_computer="yes")$ = 0,079 × 0,643 = 0,050

   $P(X \mid buys\_computer="no")\ P(buys\_computer="no")$ = 0,019 × 0,357 = 0,007

5. jika dilihat dari nilai yang diperoleh pada perhitungan di atas, diketahui bahwa nilai *buys_computer*="*yes*" lebih dari nilai *buys_computer*="*no*". Maka diperoleh kesimpulan pada Tabel 2.3 berikut

Tabel 2.3 Hasil Prediksi Data

id	*Age*	*Income*	*student*	*credit_rating*	*buys_computer*
15	≤ 30	*Medium*	*Yes*	*Fair*	*Yes*

Tabel 2.3 menyatakan bahwa salah satu pelanggan dengan umur dibawah 30 tahun, pendapatan menengah, seorang pelajar, dan tingkat kredit cukup diprediksi akan membeli computer.

Contoh di atas menggambarkan kinerja klasifikasi *naïve bayes* yaitu memprediksi data baru menggunakan data lama yang sudah diketahui. Pada penelitian ini, data yang digunakan sebagai data prediksi adalah data siswa angkatan 2015 dan data yang digunakan untuk memprediksi diterimanya siswa di perguruan tinggi negeri adalah data siswa angkatan 2010, 2011, 2013, dan 2014.

E. Perguruan Tinggi Negeri

Perguruan tinggi adalah satuan yang menyelenggarakan pendidikan tinggi dan dapat berbentuk universitas, institut, sekolah tinggi, politeknik dan akademi (Esti Setya Rini, 2012). Menurut Fuad Ihsan (2003: 23), pendidikan tinggi adalah pendidikan yang mempersiapkan peserta didik untuk menjadi anggota masyarakat yang memiliki tingkat kemampuan tinggi yang bersifat akademik dan atau professional sehingga dapat menerapkan, mengembangkan dan atau menciptakan ilmu pengetahuan, teknologi dan seni dalam rangka pembangunan nasional dan meningkatkan kesejahteraan manusia. Jadi, secara umum perguruan tinggi negeri adalah jenjang pendidikan yang lebih tinggi dari pendidikan sekolah menegah atas yang berada di bawah instansi pemerintah.

Tabel 2.4 berikut adalah jumlah perguruan tinggi negeri di Indonesia pada setiap wilayah.

Tabel 2.4 Daftar Jumlah PTN di Indonesia

Wilayah PTN	**Jumlah PTN**
Wilayah I	30
Wilayah II	10
Wilayah III	20
Wilayah IV	14
Total	**74**

Berdasarkan Tabel 2.4, diketahui bahwa jumlah perguruan tinggi negeri di Indonesia berjumlah 74 universitas, diantaranya adalah perguruan tinggi negeri di Yogyakarta yaitu Universitas Gadjah Mada (UGM), Universitas Negeri Yogyakarta (UNY), Universitas Islam Negeri (UIN), Universitas Pembangunan

Masuk Perguruan Tinggi Negeri (SNMPTN) yaitu melalui seleksi nilai rapor, Seleksi Bersama Masuk Perguruan Tinggi Negeri (SBMPTN) yaitu melalui seleksi ujian tulis yang serentak dilaksanakan di Indonesia dan Ujian Mandiri yaitu melalui seleksi ujian tulis yang jadwal pelaksanaannya ditentukan oleh masing-masing perguruan tinggi negeri.

F. Faktor-Faktor yang Berpengaruh dalam Diterimanya Siswa di Perguruan Tinggi Negeri

Keputusan seleksi masuk perguruan tinggi negeri ditentukan oleh beberapa faktor. Faktor tersebut digunakan sebagai penentu siswa dapat diterima oleh perguruan tinggi yang diinginkan. Pada penelitian ini, digunakan beberapa faktor pendukung yang dapat memprediksi tingkat diterimanya siswa di perguruan tinggi negeri. Berikut adalah faktor-faktor yang digunakan dalam penelitian ini, yaitu

1. Rata-Rata Nilai Rapor

   Fungsi pokok evaluasi hasil belajar siswa secara umum adalah untuk mengukur tingkat kemajuan siswa dalam belajar, untuk menyusun rencana belajar selanjutnya dan untuk memperbaiki proses pembelajaran (Muhammad Irham dan Novan A.W., 2013: 217). Laporan evaluasi hasil belajar siswa dituliskan pada sebuah dokumen yaitu rapor. Nilai rapor ditulis berdasarkan hasil belajar siswa dalam satu semester dan ditulis pada akhir

semester. Sebagai salah satu faktor penentu, rata-rata nilai rapor yang digunakan adalah semester satu sampai semester lima.

2. Nilai Ujian Nasional

   Ujian nasional merupakan sesuatu yang diwajibkan bagi para siswa untuk persyaratan kelulusan. Hasil ujian dapat dijadikan bukti kesanggupan siswa berpikir melalui proses yang memenuhi standar kompetensi yang ditentukan dan sesuai dengan prosedur akademik (Faiz Hidayat, 2013). Nilai ujian nasional digunakan untuk faktor yang berpengaruh masuk perguruan tinggi.

3. Tingkat Pendidikan Orang Tua

   Pendidikan merupakan suatu usaha untuk mendewasakan dan memandirikan manusia melalui kegiatan yang terencana dan disadari melalui kegiatan belajar dan pembelajaran yang melibatkan siswa dan guru (Muhammad Irham dan Novan A.W., 2013 : 19). Tingkat pendidikan diukur dari pendidikan terakhir yang ditempuh dari tingkat Sekolah Dasar (SD), Sekolah Menengah Pertama (SMP), Sekolah Menengah Atas (SMA), hingga perguruan tinggi. Kondisi sosial ekonomi berpengaruh terhadap munculnya perbedaan, salah satunya yaitu tingkat pendidikan orang tua. Tingkat pendidikan orang tua berpengaruh pada perkembangan seorang anak sejak kecil hingga dewasa. Perbedaan itu terlihat pada cara pandang terhadap suatu kondisi tertentu. Menurut Esti Setya R. (2012), tingkat pendidikan orang tua akan menentukan cara orang tua dalam membimbing dan mengarahkan anaknya dalam hal pendidikan yang sedang maupun yang akan ditempuh

anaknya. Orang tua yang berwawasan luas akan mengarahkan dan membimbing anaknya untuk terus menimba ilmu melebihi orang tuanya. Semakin tinggi tingkat pendidikan orang tua maka semakin tinggi pula minat siswa untuk melanjutkan sekolah ke perguruan tinggi.

Tingkat pendidikan orang tua ini dibagi menjadi dua atribut yaitu pendidikan ayah dan pendidikan ibu.

4. Jenis Pekerjaan Orang Tua

   Pengertian pekerjaan menurut Poerwodarminta (1996: 180) adalah sesuatu yang dilakukan untuk mencari nafkah dan untuk mengubah dirinya dengan tujuan meningkatkan taraf hidup. Semakin baik jenis pekerjaan orang tua maka semakin tinggi tingkat pendapatan orang tua. Pendapatan orang tua dapat menunjang pendidikan anaknya dalam bentuk materi, baik dalam memberi fasilitas untuk kegiatan pembelajaran, biaya sekolah, dan sebagainya. Orang tua yang mempunyai tingkat pendapatan tinggi akan mendukung penuh anaknya untuk meneruskan studi ke jenjang yang lebih tinggi, dalam hal ini anak yang ingin meneruskan studinya ke perguruan tinggi negeri.

   Jenis pekerjaan orang tua ini dibagi menjadi dua atribut yaitu pekerjaan ayah dan pekerjaaan ibu.

Salah satu aspek sebagai indikator kualitas pada sekolah menengah atas adalah tingkat diterimanya siswa di perguruan tinggi negeri. Dari faktor-faktor yang berpengaruh yang sudah dijelaskan, maka data tersebut akan dihubungkan dengan tingkat diterimanya siswa di perguruan tinggi negeri.

Tabel 2.5 berikut adalah sumber data yang digunakan pada untuk memprediksi siswa diterima di perguruan tinggi negeri.

Tabel 2.5 Sumber Data

Data Faktor yang Berpengaruh		Data Diterimanya Siswa
NIS		NIS
Nilai Rapor		Diterima di PTN
Nilai Ujian Nasional		
Pendidikan Ayah		
Pendidikan Ibu		
Pekerjaan Ayah		
Pekerjaan Ibu		

Berdasarkan Tabel 2.5, atribut yang digunakan untuk memprediksi diterimanya siswa di perguruan tinggi negeri berasal dari sumber data yang berupa data faktor-faktor yang berpengaruh dan data Sekolah Menengah Atas Negeri 1 Sleman. Data yang digunakan sebagai variabel input dalam penelitian ini adalah data siswa angkatan 2010, 2011, 2013, dan 2014. Penggunaan variabel PTN sebagai *class* target. Sementara itu, data siswa yang digunakan sebagai data prediksi adalah data siswa angkatan 2015. Data siswa tersebut diolah menggunakan bantuan *software* WEKA karena data tersebut berjumlah besar sehingga akan kesulitan jika diolah secara manual.

untuk membantu proses *data mining* mulai dari proses awal memasukkan data, proses mengolah data hingga menampilkan hasil akhir dari pengolahan data tersebut (Witten, et al, 2011: 403-404).

Gambar 2.3 Tampilan Awal WEKA

Beberapa menu dalam tampilan WEKA, diantaranya yaitu

1. *Explorer*, menu ini memberikan akses untuk semua fasilitas yang menggunakan pilihan menu dan pengisian data. Pada menu ini terdapat enam sub-menu pada bagian atas *window*, sub-menu tersebut yaitu
   a. *Preprocess*, proses pemilihan dataset yang akan diolah pemilihan filter,
   b. *Classify,* terdapat berbagai macam teknik klasifikasi yang digunakan untuk mengolah data,
   c. *Cluster,* terdapat berbagai macam teknik *cluster* yang dapat digunakan untuk mengolah data,

   d. *Associate,* terdapat berbagai macam teknik *association rules* yang dapat digunakan untuk mengolah data,
   e. *Select Atribut,* proses pemilihan aspek yang mempunyai hubungan paling relevan pada data,
   f. *Visualize*, proses menampilan berbagai plot dua dimensi yang dibentuk dari pengolahan data.
2. *Experimenter*, menu ini digunakan untuk mengatur percobaan dalam skala besar, dimulai dari *running*, penyelesaian, dan menganalisis data secara statistik.
3. *Knowledge Flow*, pada tampilan menu ini, pengguna memilih komponen WEKA dari *tool bar* untuk memproses dan menganalisis data serta memberikan alternatif pada menu *Explorer* untuk kondisi aliran data yang melewati sistem. Selain itu, *Knowledge Flow* juga berfungsi untuk memberikan model dan pengaturan untuk mengolahan data yang tidak bisa dilakukan oleh *Explorer*.
4. *Simple CLI*, menu yang menggunakan tampilan *command-line.* Menu ini menggunakan tampilan *command-line* untuk menjalankan *class* di *weka*.jar, dimana langkah pertama variabel *Classpath* dijelaskan di file *Readme*.

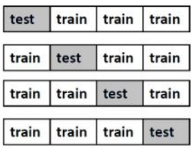
Pengolahan data dengan klasifikasi pada WEKA terdapat beberapa pilihan evaluasi yaitu *use training set, supplied test set, cross-validation*, dan *percentage split*. Pada pengolahan data siswa Sekolah Menengah Atas Negeri 1 Sleman digunakan *cross-validation*. *Cross-validation* lebih sering digunakan karena menunjukkan hasil yang maksimal untuk data yang berjumlah besar.

H. *K-Fold Cross Validation*

Menurut Witten, et al (2011: 153), *cross validation* adalah bentuk sederhana dari teknik statistik. Jumlah *fold* standar untuk memprediksi tingkat *error* dari data adalah dengan menggunakan *10-fold cross validation*.

Data yang digunakan dibagi secara acak ke dalam *k* subset yaitu D_1, D_3 ..., D_k dengan ukuran yang sama. Dataset akan dibagi menjadi data *training* dan data *testing*. Proses *training* dan *testing* dilakukan sebanyak *k* kali secara berulang-ulang. Pada iterasi ke-*i*, partisi D_i disajikan sebagai data *testing* dan partisi sisanya digunakan secara bersamaan dan berurutan sebagai data *training*. Pada iterasi pertama, subset D_2, D_3, …, D_k secara berurutan disajikan sebagai data *training* yang akan dites pada D_1. Iterasi kedua, subset D_1, D_3, …, D_k akan dites pada D_2, dan selanjutnya hingga D_k (Han, et al, 2012: 364).

Gambar 2.5 berikut adalah contoh ilustrasi 4-*fold cross validation*



Gambar 2.4 Ilustrasi *4-Fold Cross Validation*

Berdasarkan Gambar 2.4, ditunjukkan bahwa nilai *fold* yang digunakan adalah 4-*fold cross validation.* Berikut diberikan langkah-langkah pengujian data dengan 4-*fold cross validation*

a. dataset yang digunakan dibagi menjadi 4 bagian, yaitu D_1, D_2, D_3, dan D_4,

   b. D_i ($i = 1,2,3,4$) digunakan sebagai data *testing* dan dataset lainnya sebagai data *training,*
   c. tingkat akurasi dihitung pada setiap iterasi (iterasi-1, iterasi-2, iterasi-3, iterasi-4),
   d. Menghitung rata-rata tingkat akurasi dari seluruh iterasi untuk mendapatkan tingkat akurasi data keseluruhan.

I. Penelitian yang Relevan

Penelitian tentang *data mining* dengan menggunakan berbagai algoritma untuk menyelesaikan masalah di bidang pendidikan maupun bidang lainnya sudah banyak dilakukan. Beberapa diantaranya mendukung penelitian ini dengan variabel-variabel dan metode penelitian yang berkaitan.

Penelitian yang dilakukan oleh John Fredrik Ulysses dalam jurnal "*Data Mining Classification untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan dengan Metode Naïve Bayes*". Penelitian ini menggunakan algoritma *naïve bayes* dan *software RapidMiner* untuk memprediksi lama masa studi mahasiswa di STMIK Palangkaraya jurusan D3 Manajemen Informatika tahun kelulusan 2006-2008 berdasarkan jalur penerimaan mahasiswa. Klasifikasi Bayes dipilih karena telah memperlihatkan keakurasian yang tinggi dan kecepatan yang baik. Penelitian ini menunjukkan bahwa hasil klasifikasi metode *naïve bayes* dengan menggunakan 57 dataset alumni mahasiswa diperoleh 99% mahasiswa melalui jalur khusus lulus dengan waktu 5 semester, sedangkan untuk jalur SPMB 100% lulus dengan waktu di atas 5 semester. Artinya, mahasiswa yang masuk

Mining Menggunakan Metode Decision Tree C4.5". Penelitian ini menggunakan bantuan *software* WEKA dalam membentuk model prediksi tingkat kelulusan mahasiswa Jurusan Pendidikan Matematika FMIPA di Universitas Negeri Yogyakarta melalui data masuk dan kelulusan mahasiswa. Tingkat akurasi yang dihasilkan oleh *software* WEKA menggunakan algoritma C4.5 dengan 10-*fold cross validation* pada *data testing* yang berjumlah 490 dataset adalah sebesar 61,22%. *Software* WEKA digunakan untuk membantu proses dataset yang berjumlah sangat besar.

Penelitian yang dilakukan oleh Esti Setya Rini dalam skripsi yang berjudul "*Hubungan Tingkat Pendidikan Orang Tua dan Prestasi Belajar Siswa dengan Minat Siswa Melanjutkan Studi ke Perguruan Tinggi Pada Siswa Kelas XI SMA Negeri 1 Kalasan Tahun Ajaran 2011/2012"*. Penelitian ini menggunakan pendekatan kuantitatif yang dilakukan di SMA Negeri 1 Kalasan. Hasil penelitian ini menunjukkan bahwa terdapat hubungan positif dan signifikan antara tingkat pendidikan orang tua dan prestasi belajar siswa dengan minat siswa yang melanjutkan studi ke perguruan tinggi pada siswa kelas XI SMA Negeri 1 Kalasan tahun ajaran 2011/2012. Hal tersebut menunjukkan bahwa semakin tinggi tingkat pendidikan orang tua dan prestasi belajar siswa semakin tinggi pula minat siswa melanjutkan studi ke perguruan tinggi.

Berdasarkan penelitian-penelitian tersebut terdapat persamaan dan perbedaan yang digunakan dalam penulisan skripsi ini. Persamaan dengan penelitian John Fredrik Ulysses adalah penggunaan metode yang sama yaitu algoritma *naïve bayes* karena algoritma *naïve bayes* mempunyai tingkat akurasi yang tinggi dan kecepatan yang baik untuk memproses data dalam jumlah yang besar. Persamaan dengan penelitian Dimas Dwi Angen Saputra adalah penggunaan *software* WEKA karena penggunaan *software* WEKA dinilai sangat membantu perhitungan *data mining*. Perbedaan kedua penelitian tersebut terletak pada variabel dan tujuan penelitiannya, kedua penelitian tersebut memprediksi kelulusan mahasiswa dengan menggunakan data induk mahasiswa sedangkan penelitian pada skripsi ini adalah memprediksi tingkat diterimanya siswa sekolah menengah atas di perguruan tinggi negeri dengan menggunakan data induk siswa dan data kelulusan siswa. Penelitian yang dilakukan oleh Esti Setya Rini mendukung pengambilan salah satu variabel yang berpengaruh terhadap diterimanya siswa di perguruan tinggi negeri yaitu pendidikan orang tua, namun penelitian yang dilakukan Esti Setya Rini menggunakan metode pada statistika yaitu regresi linear.

DAFTAR PUSTAKA

_____. (2015). *Daftar PTN*. Diakses dari https://sbmptn.or.id/ pada tanggal 14 Juni 2015, Pukul 11.30 WIB.

Bramer, M. (2007). *Principles of Data Mining.* London: Springer.

Dimas Dwi Angen Saputra. (2014). Model Prediksi Tingkat Kelulusan Mahasiswa dengan Teknik Data Mining Menggunakan Metode Decision Tree C4.5. *Skripsi Program Studi Matematika Jurusan Pendidikan Matematika Universitas Negeri Yogyakarta*.

Esti Setya Rini. (2012). Hubungan Tingkat Pendidikan Orang Tua dan Prestasi Belajar Siswa dengan Minat Siswa Melanjutkan Studi ke Perguruan Tinggi pada Siswa Kelas XI SMA Negeri 1 Kalasan Tahun Ajaran 2011/2012. *Skripsi Jurusan Pendidikan Akuntansi Fakultas Ekonomi Universitas Negeri Yogyakarta*.

Faiz Hidayat. (2012). Kecemasan Siswa Kelas XII Jurusan Teknik Audio Video dalam Menghadapi Ujian Nasional di SMK Ma'arif NU 1 Sumpiuh. *Jurnal Program Studi Pendidikan Teknik Elektro Fakultas Teknik Universitas Negeri Yogyakarta*.

Fuad Ihsan. (2003). *Dasar-Dasar Kependidikan.* Jakarta: Rineka Cipta.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques Third Edition.* Waltham: Morgan Kaufmann.

John Fredrik U. (2013). Data Mining Classification untuk Prediksi Lama Masa Studi Mahasiswa Berdasarkan Jalur Penerimaan dengan Metode Naive Bayes. *Jurnal Magister Teknik Informatika Universitas Atma Jaya Yogyakarta*.

Kabir, M. F., Rahman, C. M., Hossain, A., & Dahal, K. (2011). Enhanced Classification Accuracy on Naive Bayes Data Mining Models. *International Journal of Computer Applications, Volume 28, No.3*.

Kusrini, & Emha Taufiq Luthfi. (2009). *Algoritma Data Mining.* Yogyakarta: Andi.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* New Jersey: John Willey and Sons, Inc.

Muhammad Irham , & Novan Andy Wiyani. (2013). *Psikologi Pendidikan: Teori dan Aplikasi dalam Proses Pembelajaran.* Yogyakarta: Ar Ruzz Media.

Nandan Supriatna. (2009). Daya Prediksi Nilai Rapor Terhadap Prestasi Belajar Mahasiswa Jalur PMDK di FPTK Universitas Pendidikan Indonesia. *Jurnal Inovec FPTK UPI, Volume V, No.14*.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* London: Springer.

Poerwadarminta. (1996). *Kamus Umum Bahasa Indonesia.* Jakarta: Balai Pustaka.

Q, Ayinde A., M., B., A.B., A., & O.A., O. (2013). Performance Evaluation of Naive Bayes and Decision Stump Algorithms in Mining Students' Educational Data. *International Journal of Computer Science Issues, Vol. 10, Issue 4, No. 1*.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2006). *Database System Concepts Fifth Edition.* New York: Mc Grow Hill.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques Third Edition.* Burlington: Morgan Kaufmann.